ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 3, Nomor 4, Tahun 2014, Halaman 595 - 604
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

IDENTIFIKASI LAMA STUDI BERDASARKAN KARAKTERISTIK MAHASISWA MENGGUNAKAN ALGORITMA C4.5 (Studi Kasus Lulusan Fakultas Sains dan Matematika Universitas Diponegoro Tahun 2013/2014)

Bramaditya Swarasmaradhana[1], Moch. Abdul Mukid[2], Agus Rusgiyono[3]
[1]Mahasiswa Jurusan Statistika FSM UNDIP
[2,3]Staff Pengajar Jurusan Statistika FSM UNDIP

ABSTRACT

Based on academics regulation No. 209/PER/UN7/2012, the study period of students in Diponegoro University has been scheduled for 4 years. In this study the graduation status of students that graduate under or equal to 4 years categorized as graduate on time, meanwhile students that graduate over 4 years categorized as graduate out of time. Hence, it is important to understand the profile of students who graduate on time and out of time based on gender, majors, GPA, organizational experience, part time experience, scholarship, students origin and pathways scholar. The purpose of this study is to identify those students profiles using Algorithm C4.5. Algorithm C4.5 contructs a decision tree that able to handle missing values, able to handle continues attribute and able to simplify the trees by pruning. The accuration of the Algorithm C4.5 is 84.475% and the number of the nodes are 20 nodes where 13 nodes are leaf nodes. The students profile that identified graduate on time are students of Physics who had received scholarship and a woman; students of Chemistry with GPA > 3.06; students of Statistics with GPA > 3.43 from SNMPTN also PSSB and students of Mathematics with GPA > 2.96.

Keywords: Study Period, Algorithm C4.5, Decision Tree.

1. PENDAHULUAN

Pendidikan penting bagi setiap orang sebagai bekal untuk dapat melangsungkan kehidupannya. Pentingnya pendidikan bagi setiap orang di dalam sebuah negara akan memberikan pengaruh positif terhadap negara tersebut karena dengan pendidikan akan meningkatkan kualitas sumber daya manusia sehingga bagi negara tentu akan menambah daya saing terhadap negara lain. Menurut UU No. 12 tahun 2012, untuk meningkatkan daya saing bangsa dalam menghadapi globalisasi di segala bidang diperlukan pendidikan tinggi yang mampu mengembangkan ilmu pengetahuan dan teknologi serta menghasilkan intelektual, ilmuwan atau profesional yang berbudaya dan kreatif, toleran, demokratis, berkarakter tangguh, serta berani membela kebenaran untuk kepentingan bersama.

Pendidikan adalah usaha untuk mewujudkan suasana belajar dan proses pembelajaran. Perguruan tinggi adalah satuan pendidikan yang menyelenggarakan pendidikan tinggi. Menurut Djamarah (2002), untuk memperoleh hasil belajar yang baik harus melalui proses tertentu yang dipengaruhi oleh faktor dari dalam diri individu dan di luar individu. Banyak faktor dari luar anak didik yang mempengaruhi proses belajar seperti lingkungan sosial dan lingkungan alami. Sedangkan faktor dari dalam yang dapat mempengaruhi anak didik seperti halnya minat, kecerdasan, dan motivasi, dimana motivasi dalam perguruan tinggi salah satunya berupa ijazah kelulusan.

Berbeda dengan jenjang pendidikan dasar dan menengah, pada tingkat perguruan tinggi terutama program sarjana S1 memiliki syarat kelulusan bagi setiap mahasiswa adalah telah menempuh minimal 144 sampai 160 SKS. Sedangkan untuk lama studi mahasiswa, dikarenakan penelitian ini mengambil studi kasus di Universitas Diponegoro maka dengan berpedoman pada peraturan akademik Universitas Diponegoro No.209/PER/UN7/2012, lama studi mahasiswa ditetapkan dapat ditempuh dalam kurun waktu 4 tahun atau 8 semester dengan batas maksimal adalah 7 tahun atau 14 semester.

Berdasarkan uraian tersebut penelitian ini membahas mengenai identifikasi lama studi mahasiswa menggunakan Algoritma C4.5. Algoritma C4.5 merupakan salah satu metode klasifikasi untuk mengkonstruksikan pohon keputusan (*decision trees*) yang terdapat pada data mining. Algoritma C4.5 merupakan pembaharuan metode ID3 oleh Quinlan. Kelebihan Algoritma C4.5 dari metode pohon keputusan sejenis adalah bahwa algoritma C4.5 mampu mengatasi atribut yang bersifat kontinu, mengatasi nilai yang hilang dan melakukan pemangkasan pohon yang kompleks. Dalam penelitian ini peneliti ingin membentuk pohon klasifikasi untuk mengidentifikasi mahasiswa yang lulus dengan lama studi $\leq$ 4 tahun yang akan dikategorikan tepat waktu dan mahasiswa yang lulus dengan lama studi $>$ 4 tahun yang dikategorikan tidak tepat waktu berdasarkan faktor jenis kelamin, jurusan, IPK, beasiswa, pengalaman berorganisasi, kerja paruh waktu, daerah asal dan jalur masuk. Penelitian ini mengambil sampel lulusan Fakultas Sains dan Matematika Universitas Diponegoro periode April 2013 sampai dengan Januari 2014. Hasil dari penelitian ini berupa pohon keputusan mengenai lama studi mahasiswa baik yang lulus tepat waktu dan mahasiswa yang lulus tidak tepat waktu.

2. TINJAUAN PUSTAKA

2.1. Pendidikan Tinggi

Menurut Undang-Undang nomor 12 tahun 2012, definisi pendidikan tinggi adalah jenjang pendidikan setelah pendidikan menengah yang mencakup program diploma, program sarjana, program magister, program doktor dan program profesi serta program spesialis, yang diselenggarakan oleh perguruan tinggi berdasarkan kebudayaan bangsa Indonesia kemudian definisi mahasiswa adalah peserta didik pada jenjang pendidikan tinggi. Menurut peraturan akademik Universitas Diponegoro No. 209/PER/UN7/2012, memiliki beban studi sekurang-kurangnya 144 sks dan sebanyak-banyaknya 160 sks yang dijadwalkan untuk 8 semester dan dapat ditempuh dalam waktu kurang dari 8 semester dan paling lama 14 semester.

2.2. Motivasi dalam Belajar

Menurut Djamarah (2002), definisi belajar adalah serangkaian kegiatan jiwa raga untuk memperoleh suatu perubahan tingkah laku sebagai hasil dari pengalaman individu dalam interaksi dengan lingkungannya yang menyangkut kognitif, afektif dan psikomotor. Banyak unsur yang mempengaruhi belajar dan hasil belajar setiap anak didik. Unsur-unsur tersebut terbagi menjadi unsur dari dalam maupun dari luar individu. Unsur dari luar meliputi faktor lingkungan yaitu lingkungan alami di sekitar tempat tinggal anak didik dan lingkungan sosial budaya dimana setiap anak didik berinteraksi.

2.3. Algoritma C4.5

Algoritma C4.5 adalah algoritma pembentukan pohon keputusan yang merupakan pengembangan dari algoritma pohon keputusan ID3 (*Iterative Dichotomiser 3*) yang diciptakan oleh J. Ross Quinlan pada 1993. Menurut Witten *et al.* (2011), Algoritma C4.5 memiliki keunggulan dibandingkan dengan ID3 yaitu mampu mengatasi nilai yang hilang (*missing value*), mengatasi data bertipe kontinu dan melakukan pemangkasan pohon (*pruning trees*). Dalam konstruksi pohon keputusan data terbagi menjadi sampel pelatihan (*training sample*) dan sampel pengujian (*testing sample*). Sampel pelatihan digunakan untuk mengkonstruksikan pohon keputusan dan sampel pengujian digunakan untuk menguji hasil konstruksi pohon.

2.3.1. Pembentukan Pohon Keputusan Algoritma C4.5

Menurut Ruggieri (2002), algoritma pembentukan pohon keputusan menggunakan sampel pelatihan yang terdiri atas kumpulan kasus dimana setiap kasus memiliki atribut dan kelas. Atribut-atribut dalam kasus dapat bertipe kontinu maupun diskret tetapi pada kelas setiap kasus harus bertipe diskret yang dapat dinotasikan $C_1,...,C_n$.

Pohon Keputusan adalah pohon data yang terdiri atas simpul keputusan dan simpul daun yang terstruktur. Simpul daun tersebut memiliki kelas-kelas dan simpul keputusan menguji beberapa atribut sampai diperoleh atribut yang terpilih sebagai pemilah. Setiap cabang pada pengujian tersebut menghasilkan simpul anak (*Child Node*) di bawah simpul induknya (*Parent Node*).

Pembentukan pohon keputusan menggunakan prinsip membagi sampel pelatihan menjadi beberapa sub-himpunan yang berbeda. Langkah awal dari konstruksi pohon keputusan algorithma C4.5 adalah mencari simpul akar. Berikut ini langkah-langkah konstruksi pohon keputusan menggunakan Algoritma C4.5 berdasarkan Ruggieri (2002):

1. Misalkan T adalah himpunan kasus-kasus yang akan dibuat simpul dimana kasus-kasus tersebut memiliki kelas dan atribut-atribut. Frekuensi terboboti $freq(C_j,T)$ diperoleh dari perhitungan T dan kelas yang dihasilkan adalah C_j, untuk setiap $j \in \{1,2,3,...,n\}$
2. Jika semua kasus berada dalam kelas C_j yang sama maka simpul yang dihasilkan adalah simpul daun yang dilabeli dengan kelas C_j sebagai kelas terbanyak. Kesalahan klasifikasi pada simpul daun merupakan kasus-kasus dalam T yang berbeda kelas dengan kelas C_j.
3. Jika T berisi kasus yang memiliki dua atau lebih kelas maka dapat dihitung *information gain* dari setiap atribut tersebut. Untuk atribut diskret, *information gain* disesuaikan dengan pembagi dalam T dengan nilai atribut yang sudah diketahui sebelumnya. Untuk atribut kontinu, *information gain* disesuaikan dengan pembagi T ke dalam dua irisan (biner) yang dilabeli kasus dengan nilai atribut kurang dari atau sama dengan nilai ambang batas ($A \leq v$) dan nilai atribut dengan nilai atribut lebih besar dari nilai ambang batas ($A > v$).
4. Atribut dengan nilai *information gain* tertinggi terpilih sebagai pemilah dalam simpul tersebut.

5. Simpul keputusan memiliki cabang sebanyak s yaitu $T_1, \ldots, T_s$ dimana $s = 2$ untuk atribut kontinu dan $s = h$ untuk atribut diskret dengan nilai h yang sudah diketahui.
6. Untuk setiap $i = \{1,2,\ldots,s\}$, jika T_i tidak memiliki cabang lagi maka simpul tersebut secara langsung menjadi simpul daun yang diberi label kelas terbanyak di bawah simpul induknya dan kesalahan klasifikasi bernilai 0.
7. Apabila T_i memiliki cabang lagi maka pemilahan diproses kembali menggunakan kasus-kasus dalam T_i. Catatan khusus untuk kasus-kasus dengan nilai yang hilang pada atribut terpilih tersebut dilakukan proses pemilihan pemilah pada setiap simpul anaknya dengan pembobotan banyak kasus yang diketahui dibagi dengan banyak kasus pada simpul tersebut.
8. Terakhir, kesalahan klasifikasi simpul dihitung dari penjumlahan dari kesalahan-kesalahan simpul anak yang dibandingkan dengan simpul induknya.

2.3.2. Prosedur Pemilahan Algoritma C4.5

Information gain dari sebuah atribut a dari himpunan T dihitung jika a adalah diskret dan $T_1, \ldots\ldots, T_s$ adalah sub-himpunan dari T yang terdiri dari kasus-kasus yang nilainya sudah diketahui. Untuk mendapatkan *information gain* dari atribut a atau *gain(a)* dibutuhkan *entropy* keseluruhan kelas pada himpunan T atau *info(T)* dan *entropy* masing-masing atribut pada himpunan T atau *info(T_i).* Rumus dari *Gain(a)* adalah sebagai berikut:

$$Gain(a) \quad = info(T) - \sum_{i=1}^{s} \frac{|T_i|}{|T|} \times info(T_i) \qquad (1)$$

dimana

$$info(T) \quad = -\sum_{j=1}^{n} \frac{freq(C_j,T)}{|T|} \times {}^2\log\left(\frac{freq(C_j,T)}{|T|}\right) \qquad (2)$$

nilai *entropy* untuk setiap atribut *i,*

$$info(T_i) \quad = -\sum_{j=1}^{n} \frac{freq(C_j,T_i)}{|T_i|} \times {}^2\log\left(\frac{freq(C_j,T_i)}{|T_i|}\right) \qquad (3)$$

keterangan:

$|T|$ = Banyak kasus dalam himpunan T
$|T_i|$ = Banyak kasus dalam sub-himpunan T_i
$Freq(C_j,T)$ = Banyak dari kasus-kasus dalam himpunan T yang memiliki kelas C_j

Fungsi *info(T)* pada Persamaan (2) adalah fungsi *entropy* dimana *entropy* menurut Quinlan (1993) adalah rataan jumlah informasi yang dibutuhkan untuk mengidentifikasi kelas suatu kasus ke dalam himpunan T. Nilai dari setiap penghitungan *entropy* memiliki satuan *bits* atau *binary digits.*

Jika a adalah atribut kontinu maka kasus dalam T dengan nilai atribut tersebut diurutkan dari yang terkecil sampai terbesar. Dimisalkan nilai hasil pengurutan adalah $w_i, \ldots\ldots, w_m$, untuk $i \in \{1,2,3,\ldots,m\text{-}1\}$ dimana nilai $v=\frac{(w_i+w_{i+1})}{2}$ dan pemisahan yang terjadi untuk atribut bertipe kontinu adalah:

$$T_1 \; = \{w_j \mid w_j \leq v\} \text{ dan } T_1 \; = \{w_j \mid w_j > v\} \qquad (4)$$

Untuk setiap nilai v, nilai *gain* dihitung dengan mempertimbangkan prosedur pemisahan (4). *Information gain* untuk a didefinisikan sebagai nilai

maksimum *gain* dari *v* dimana nilai *v* merupakan nilai ambang batas untuk atribut kontinu.

2.3.3. Pemangkasan Pohon Keputusan

Menurut Ning Tan *et al.* (2006), metode pemangkasan ini menggunakan batas atas kepercayaan distribusi binomial dengan nilai α untuk Algoritma C4.5 yaitu α = 25% dimana *"E"* adalah banyaknya kesalahan klasifikasi, *"N"* adalah banyaknya kasus pada simpul dan *"f"* adalah rasio kesalahan pada suatu simpul yaitu $f = E/N$. Proses pemangkasan terjadi apabila kesalahan terprediksi (*predicted error*) pada simpul anak lebih besar daripada simpul induk. Penghitungan kesalahan terprediksi merupakan hasil dari penjumlahan $e \times N$ tiap cabang daun dimana *e* atau batas atas kepercayaan menggunakan rumus sebagai berikut:

$$e = \frac{f+\frac{z^2_{\alpha/2}}{N}+z_{\alpha/2}\sqrt{\frac{f}{N}-\frac{f^2}{N}+\frac{Z^2_{\alpha/2}}{4N^2}}}{1+\frac{z^2_{\alpha/2}}{N}} \qquad (5)$$

2.3.4. Pengukuran Ketepatan Hasil Klasifikasi

Menurut Han *et al.* (2012), hasil klasifikasi algoritma pohon keputusan dapat diuji dengan menggunakan Matriks Konfusi. Matriks ini terdiri atas jumlah kasus yang diklasifikasikan secara tepat dan tidak tepat. Metode ini menggunakan level matriks seperti pada Tabel 1. berikut ini :

Tabel 1. Matriks Konfusi Tiga Kelas

Aktual	Prediksi (+)	Prediksi (-)	*Total*
(+)	A	B	*P*
(-)	C	D	*N*
Total	*P'*	*N'*	*P+N*

Untuk mengetahui hasil kinerja dari klasifikasi yang dikonstruksikan dapat diukur menggunakan akurasi. Akurasi adalah prosentase dari keseluruhan set yang diklasifikasikan secara tepat. Rumusnya:

$$\text{Akurasi} = \frac{\text{A+D}}{A+B+C+D} \text{x } 100\%$$

3. METODOLOGI PENELITIAN

3.1. Sumber Data

Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari buku kelulusan mahasiswa dan data primer yang diperoleh dengan melakukan survey alumni Universitas Diponegoro Fakultas Sains dan Matematika periode April 2013, Agustus 2013. Oktober 2013 dan Januari 2014. Data Alumni tersebut berjumlah 351 orang kemudian diambil 200 orang sebagai sampel.

3.2. Atribut Penelitian

Atribut yang digunakan dalam penelitian ini adalah lama studi (tepat waktu dan tidak tepat waktu), IPK, jenis kelamin (pria dan wanita), jurusan (Matematika, Statistika, Informatika, Fisika, Biologi dan Kimia), beasiswa (pernah dan tidak pernah), organisasi (pernah dan tidak pernah), kerja paruh waktu (pernah dan

tidak pernah), daerah asal (Semarang dan luar Semarang) dan terakhir jalur masuk (SMMPTN, UM dan PSSB).

3.3. Metode Penelitian

Dalam penelitian ini metode yang digunakan adalah studi literatur yang diperoleh dari sumber-sumber resmi, seperti perpustakaan serta situs-situs di internet. Studi kasus ini juga akan didukung dengan pemrograman pembentukan pohon keputusan C4.5 dari data dengan menggunakan *software* WEKA dan Microsoft Excel. Pengolahan klasifikasi didalam *software* WEKA menggunakan J48 yang merupakan nama aplikasi Algoritma C4.5 pada *Software* WEKA. J48 secara komprehensif dapat menunjukkan konstruksi pohon keputusan algoritma C4.5 serta dapat menunjukkan akurasi klasifikasi. Adapun langkah-langkah untuk mencapai tujuan penelitian ini adalah sebagai berikut:

1. Membagi data menjadi sampel pelatihan dan sampel pengujian dengan prosentase masing-masing sampel pelatihan 80% dan sampel pengujian 20% menggunakan pengambilan sampel acak sederhana.
2. Mengkonstruksikan pohon keputusan Algoritma C4.5 menggunakan sampel pelatihan.
3. Melakukan analisis terhadap hasil pohon keputusan yang terbentuk dan menghitung nilai akurasi pohon.
4. Melakukan pemangkasan pohon dan menghitung nilai akurasi pohon.
5. Mengidentifikasikan profil mahasiswa yang lulus tepat waktu dan lulus tidak tepat waktu.
6. Menguji pohon keputusan menggunakan sampel pengujian.

4. PEMBAHASAN

4.1 Algoritma C4.5

4.1.1. Konstruksi Pohon Keputusan

Konstruksi pohon keputusan menghasilkan pohon dengan banyak simpul mencapai 51 simpul yang terdiri dari simpul akar, simpul keputusan dan simpul daun. Banyak simpul daun sendiri mencapai 31 simpul yang dilabeli dengan kelasnya masing-masing. Simpul akar merupakan simpul yang terletak paling atas. Dalam penelitian ini atribut jurusan terpilih sebagai pemilah pada simpul akar. Atribut Jurusan terpilih karena nilai *information gain* terbesar diantara atribut lainnya. Berikut ini proses penghitungan manual mendapatkan *information gain* atribut jurusan yang menjadi pemilah pada simpul akar:

1. Hitung frekuensi tiap kelas (tepat waktu dan tidak tepat waktu)

Kelas	Frek($Cj,\|T\|$)
Tepat Waktu	61
Tidak Tepat Waktu	99
Total ($\|T\|$)	160

2. Hitung nilai *entropy* kelas yang disimbolkan *Info(T)*

$$info(T) = -\sum_{j=1}^{n} \frac{freq(C_j,T)}{|T|} \times {}^{2}\log\left(\frac{freq(C_j,T)}{|T|}\right)$$

$$= -\sum_{j=1}^{2} \frac{freq(C_j,T)}{|T|} \times {}^{2}\log\left(\frac{freq(C_j,T)}{|T|}\right)$$

$= -\left(\frac{61}{160} \times {}^{2}log\left(\frac{61}{160}\right) + \frac{99}{160} \times {}^{2}log\left(\frac{99}{160}\right)\right)$

$= -\ (0{,}38 \times {}^{2}\log 0{,}38 + 0{,}62\ {}^{2}\log 0{,}62)$

$= -\big((-0{,}53) + (-0{,}43)\big)$

info(T) = 0,96 bits

3. Hitung frekuensi masing-masing kategori pada atribut jurusan berdasarkan kelasnya.

Jurusan	Frekuensi		Total
	Tepat Waktu	Tidak Tepat Waktu	
Matematika	19	8	27
Statistika	10	23	33
Informatika	0	18	18
Biologi	12	23	35
Kimia	14	14	28
Fisika	6	13	19
Total	80	120	160

4. Hitung nilai *Information Gain* yang disimbolkan *Gain* (Jurusan)

Gain (Jurusan) $= info(T) - \sum_{i=1}^{6} \frac{|T_i|}{|T|} \times info(T_i)$

$info(T)$ = 0,96 *bits*

info(T_i) $= -\sum_{j=1}^{n} \frac{freq(C_j,Ti)}{|Ti|} \times {}^{2}log\left(\frac{freq(C_j,Ti)}{|Ti|}\right)$

Info(Matematika) $= -\left(\frac{19}{27}\ {}^{2}log\frac{19}{27} + \frac{8}{27}\ {}^{2}log\frac{8}{27}\right)$

$= 0{,}877$

Info(Statistika) $= -\left(\frac{10}{33}\ {}^{2}log\frac{10}{33} + \frac{23}{33}\ {}^{2}log\frac{23}{33}\right)$

$= 0{,}885$

Info(Informatika) $= -\left(\frac{0}{18}\ {}^{2}log 0 + \frac{18}{18}\ {}^{2}log\frac{18}{18}\right)$

$= 0{,}000$

Info(Biologi) $= -\left(\frac{12}{35}\ {}^{2}log\frac{12}{35} + \frac{23}{35}\ {}^{2}log\frac{23}{35}\right)$

$= 0{,}928$

Info(Kimia) $= -\left(\frac{14}{28}\ {}^{2}log\frac{14}{28} + \frac{14}{28}\ {}^{2}log\frac{14}{28}\right)$

$= 1{,}000$

Info(Fisika) $= -\left(\frac{6}{19}\ {}^{2}log\frac{6}{19} + \frac{13}{19}\ {}^{2}log\frac{13}{19}\right)$

$= 0{,}900$

Gain (Jurusan) $= 0{,}96 - \left(\frac{27}{160} \times 0{,}877 + \frac{33}{160} \times 0{,}885 + \frac{18}{160} \times 0{,}000 + \frac{35}{160} \times 0{,}928 + \frac{28}{160} \times 1{,}000 + \frac{19}{160} \times 0{,}900\right)$

Gain (Jurusan) = (0,96 – 0,82) *bits*

Gain (Jurusan) = 0,14 *bits*

Nilai *information gain* atribut jurusan sebesar 0,14 merupakan yang tertinggi diantara atribut lainnya. Sehingga atribut jurusan terpilih sebagai pemilah dalam simpul akar. Proses selanjutnya yaitu melakukan pembentukan

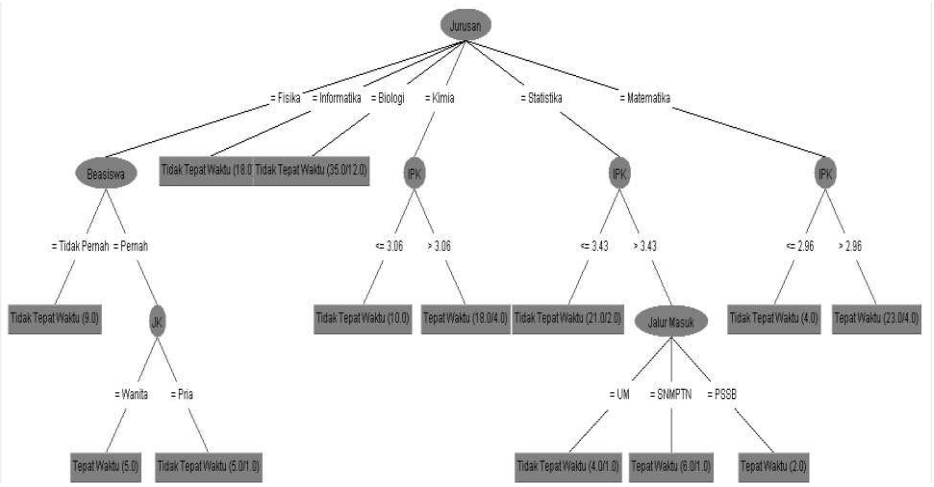
pohon sesuai dengan langkah-langkah pembentukan simpul akar. Kasus-kasus dipilah sesuai dengan sub-himpunan atribut terpilih untuk dilakukan konstruksi pohon. Dengan nilai *info(T)* merupakan nilai *entropy* simpul keputusan dan nilai *info(Ti)* merupakan nilai masing-masing atribut. Pada konstruksi pohon keputusan menggunakan Algoritma C4.5 ini didapatkan nilai akurasi klasfikasi mencapai 90%.

Algoritma C4.5 dapat mengatasi atribut bertipe kontinu seperti atribut IPK yang menjadi pemilah pada simpul keputusan cabang Jurusan Statistika. Pemisahan pada atribut bertipe kontinu akan menghasilkan cabang biner yaitu $T_1 = \{w_j \mid w_j \leq v\}$ dan $T_2 = \{w \mid w_j > v\}$.

Proses perhitungan pada atribut bertipe kontinu seperti IPK pada Jurusan Statistika adalah dengan mencari nilai *entropy* dari setiap kemungkinan nilai ambang batas yang ada dengan terlebih dahulu mengurutkan data dari yang terkecil hingga yang terbesar kemudian mencari nilai ambang batas. Rumus mencari nilai ambang batas v adalah $v = \frac{w_i + w_{i+1}}{2}$. Pada atribut IPK tersebut terpilih nilai ambang batas v = 3,43, sehingga pemisahan pemilah IPK adalah IPK $\leq$ 3,43 dan IPK> 3,43.

4.1.2. Pemangkasan Pohon Keputusan

Pemangkasan (*pruning*) dilakukan untuk mengurangi kompleksitas pohon agar menjadi lebih sederhana. Dengan pemangkasan jumlah simpul akan menjadi berkurang sehingga jumlah simpul daun juga akan berkurang. Pemangkasan dilakukan menggunakan operasi penggantian sub-pohon (*subtree replacement*). Pemangkasan dilakukan dengan mengestimasi kesalahan terprediksi setiap simpul melalui pendekatan statistik menggunakan batas atas kepercayaan distribusi binomial seperti pada Persamaan (5). Simpul akan dipangkas apabila kesalahan terprediksi simpul daun lebih besar daripada simpul keputusannya. Berikut ini adalah pohon keputusan menggunakan Algoritma C4.5 setelah pemangkasan:



Konstruksi Pohon Keputusan Algoritma C4.5

Konstruksi pohon keputusan setelah pemangkasan menghasilkan akurasi sebesar 84,475%. Konstruksi pohon keputusan juga menunjukkan banyak simpul yang terbentuk menjadi 20 simpul dan jumlah simpul daun menjadi 13

simpul. Banyaknya simpul daun menunjukkan banyaknya profil dengan masing-masing kelas pada simpul daun. Berikut ini profil mahasiswa yang lulus tepat waktu dan lulus tidak tepat waktu:

a. Lulus Tepat Waktu (lama studi $\leq$ 4 tahun)
   1. Mahasiswa Jurusan Fisika yang pernah mendapatkan beasiswa dan berjenis kelamin wanita.
   2. Mahasiswa Jurusan Kimia dengan IPK lebih dari 3,06.
   3. Mahasiswa Jurusan Statistika dengan IPK > 3,43 dan masuk universitas melalui jalur SNMPTN dan PSSB.
   4. Mahasiswa Jurusan Matematika dengan IPK > 2,96.

b. Lulus Tidak Tepat Waktu (lama studi > 4 tahun)
   1. Mahasiswa Jurusan Fisika yang tidak pernah mendapatkan beasiswa.
   2. Mahasiswa Jurusan Fisika yang pernah mendapatkan beasiswa dan berjenis kelamin pria.
   3. Mahasiswa Jurusan Informatika.
   4. Mahasiswa Jurusan Biologi.
   5. Mahasiswa Jurusan Kimia dengan IPK $\leq$ 3,06.
   6. Mahasiswa Jurusan Statistika dengan IPK $\leq$ 3,43.
   7. Mahasiswa Jurusan Statistika dengan IPK > 3,43 dan masuk universitas melalui Ujian Mandiri (UM).
   8. Mahasiswa Jurusan Matematika dengan IPK $\leq$ 2,96.

Setelah didapatkan hasil konstruksi pohon dengan nilai akurasi mencapai 84,375%, maka untuk mengetahui apakah hasil konstruksi pohon baik digunakan untuk memprediksi kemungkinan kelas pada kasus-kasus selanjutnya, pohon konstruksi Algoritma C4.5 tersebut diujikan dengan memasukkan data sampel pelatihan kedalam pohon konstruksi. Pada penelitian ini dengan menggunakan 20% dari sampel berukuran 200 sebagai sampel pengujian, diperoleh nilai akurasi pada sampel pengujian mencapai 67,5%.

5. KESIMPULAN

Setelah melakukan pengklasifikasian dengan konstruksi pohon keputusan menggunakan Algoritma C4.5 pada lama studi mahasiswa Fakultas Sains dan Matematika periode tahun 2013/2014, dapat diambil kesimpulan sebagai berikut:

1. Konstruksi pohon keputusan yang terbentuk menggunakan Algoritma C4.5 menghasilkan pohon dengan banyak simpul mencapai 51 simpul dimana 31 diantaranya adalah simpul daun dan atribut jurusan terpilih sebagai simpul akar dalam pembuatan pohon keputusan. Setelah dilakukan pemangkasan pohon jumlah simpul menjadi 20 simpul dan simpul daun menjadi 13 simpul.
2. Berdasarkan pengukuran kinerja klasifikasi menunjukkan bahwa akurasi atau ukuran ketepatan klasifikasi mencapai 90 %. Setelah dilakukan pemangkasan akurasi menjadi 84,375 %. Berdasarkan pengukuran akurasi hasil klasifikasi Algoritma C4.5 menggunakan sampel pengujian yang berjumlah 40 sampel menunjukkan akurasi sebesar 67,5 %.
3. Identifikasi kelulusan mahasiswa yang lulus tepat waktu (lama studi $\leq$ 4 tahun) menghasilkan 5 profil yaitu :
   a. Mahasiswa jurusan Fisika yang pernah mendapatkan beasiswa dan berjenis kelamin wanita.

b. Mahasiswa jurusan Kimia dengan IPK > 3,06.
c. Mahasiswa jurusan Statistika dengan IPK > 3,43 dan masuk universitas melalui jalur SNMPTN dan jalur PSSB.
d. Mahasiswa jurusan Matematika dengan IPK > 2,96.

DAFTAR PUSTAKA

Djamarah, S.B. 2002. Psikologi Belajar. Rineka Cipta. Jakarta.

Han, J, Kamber, M and Pei, J. 2012. *Data Mining Concepts and Technique*. Third Edition. Elsevier, Inc. *Massachusetts.*

Quinlan, J.R. 1993. *C4.5 : Programs For Machine Learning.* Morgan Kaufmann Publisher, Inc. San Mateo.

Ruggieri, S. 2002. *Efficient C4.*5.(http://www.di.unipi.it/~ruggieri/Papers/ec45.pdf, diakses pada tanggal 02 Maret 2014).

Rokach, L and Maimon, O. 2008. *Data Mining With Decision Trees : Theory and Applications*. World Scientific Publishing Co. Pte. Ltd. Singapura.

Tan, P, Steinbach, M and Kupar V. 2006. *Introduction to Data Mining*. Addison-Wesley. Boston.

Witten, I. Frank, H.E and Hall, M.A. 2011. *Data Mining : Practical Machine Learning Tools and Technique.* Third Edition. Elsevier, Inc. . *Massachusetts*.